Edge Computing

Abhinav Mahajan of International Institue of Information Technology, Bangalore

***Abstract*—This paper explains and summarizes edge computing, its invention, characteristics, promise and future scope.**

***Index Terms*—Edge Computing, Artificial Intelligence (AI), IoT, data centers, Cloud, Cloudlet, Internet, Virtual Reality(VR), latency, bandwidth, nodes, network core and edge.**

I. Introduction

A. What is edge computing?

The gist of edge computing is to offload computation or cyber foraging to a nearby dispersed cloud device(cloudlet) which is in closest proximity to the user. [1]

Cloudlets, micro-data centers or fog nodes will act as storage and computational nodes, just as a cloud data system does, but is decentralized and is much closer to the user (in fact it will be the first hop from the user to the internet).

This opens up a wide array of opportunities to use latency sensitive applications and use them at their desired speed and also have a better and more secure, hierarchical architecture of internet.

B. How are IoT (Internet of Things) and Edge Computing related?

IoT is the process of connecting devices to the web of internet. IoT contains the internet user's data and its flow is strictly controlled by it.

IoT will be used to connect a device to its corresponding hierarchical structured edge node system. An IoT gateway can be used to send data to the edge system, and from the edge to its centralized data center (if necessary).

IoT relies on cloud or centralized systems of internet, which means there is a reduction in operating efficiency and much slower responses.

Edge computing reduces traffic in the Internet as a centralized cloud will not be bombarded with queries (flow control) and even reduce burden on intermediate nodes (congestion control). [2]

Along with this latency and bandwidth are made better.

C. Brief chronology of the invention of Edge Computing: -

First and foremost step taken in 1990s by Akamai, when he launched a Content Delivery Network (CDNs) and local servers were provided to the internet. Edge computing is an extension of this idea and incorporates the newer features of the mordern day internet.

In 1997, the paper "agile application-aware adaptation for mobility" demonstrated how to resource manage the mobile devices computing power and offload it to powerful servers (pervasive computing) and in 2001, "Pervasive computing: vision and challenges." Summarized the concept. [10]

In 2001, P2P or peer-to-peer applications were proposed to the already existing client to server structure. This showcased how to hierarchically scale the internet.

In 2006 cloud computing started to emerge, when amazon and google began using it.

In 2009, Satyanarayanan introduced edge/cloudlet computing in a two-tier structure with the primary objective being reduced latency and the cloudlets act as a soft cache for data. [7]

In 2012, Cisco introduced the term fog computing for decentralized data center hierarchy of internet with scalability. [8]

These were the key events which led to the formation of edge computing.

Currently, Investment and research has been exponentially increasing.

ETSI ISG MEC (European Telecommunications Standards Institute Industry Specification Group for Mobile Edge Comuting) has become a reality.

Multiple Companies including Cisco, Microsoft, Google, Amazon and Dell have held and taken part in symposiums for edge computing and shown the world hands-on live deployment of prototypical model Cloudlets.

II. Current framework

Aims and key features of Edge Computing: -

A. Low Latency Response Services

Figure 2 and 3 (page 2) illustrates the difference between edge computing response time versus others.

Cloud Data Centers are usually far away, multiple hops away from the host. Cloud services therefore have a limitation, as they cannot service latency/bandwidth sensitive products. The crux of edge computing is to bring cloud services as close to the user as possible.

Notably, Mobile-only computation in the local host itself takes more time than edge computing.

Some applications like multimedia streaming, video transcoding, Virtual Reality headsets, optical recognition services, services which need feed from rapidly changing visual data, real-time cognitive assistance service et cetera are all extremely latency sensitive. Up to 16ms latency is supportable for their smooth, jitter-free functioning. While cloud services may take 300-600 milliseconds to respond, edge computation can offer up to 4 milliseconds. [1]

Edge computing is essential in this case, as we can see faraway cloud services cannot support such services.

Some cognitive recognition software's, or highly AI dominant real time assistant applications may require a lot of code and will not be wise to download such heavy software's

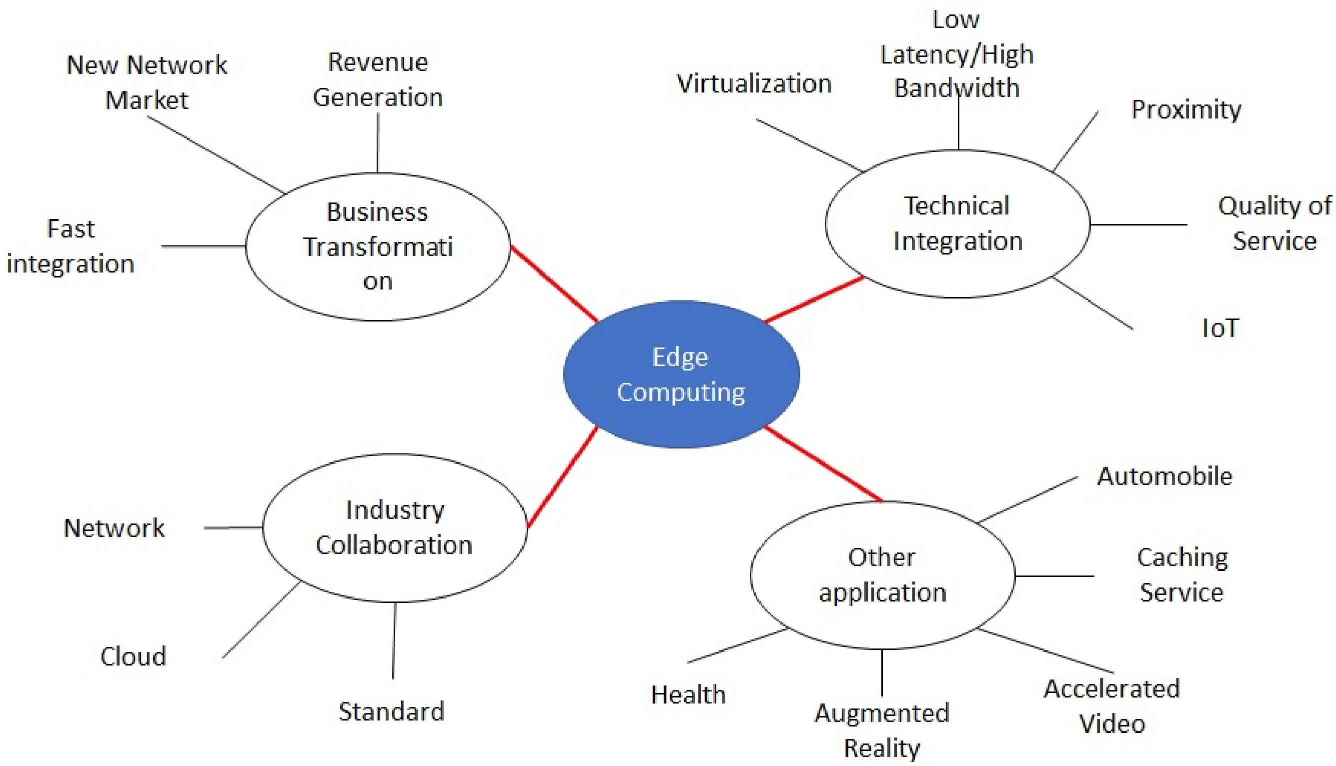

Fig. 1. Various fields of Edge Computing

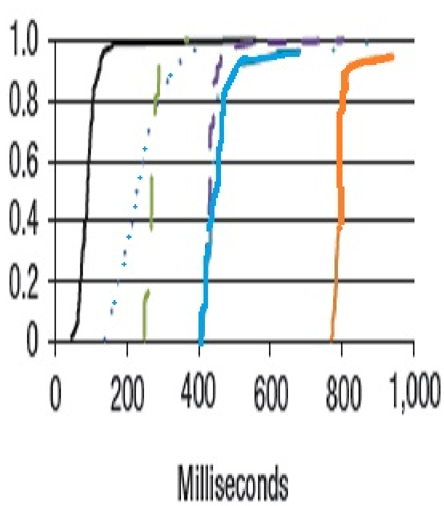

Fig. 2. Graph comparing response times of different network systems

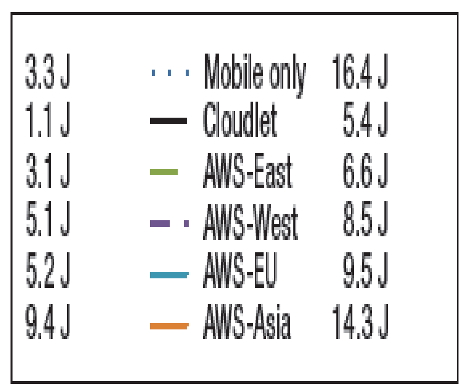

Fig. 3. Graph key

on local devices, especially phones, they may not be able to support such high computational data, and this can also result in damage of battery life as unnecessary computational burden is given to the local host when in reality it can be offloaded to a dedicated node in close proximity.

B. Hierarchical Architecture Layout of Internet

This is pivotal, this is good for data back-up, security, alternate nodes in case of node failure, congestion control in network core, reduce traffic and burden of few congested routers or main centralized data centers, and due to dedicated edge nodes for computation, Big Data can be crunched and hashed into corresponding metadata, reducing overheads, size of data and better transportation times to and from the edge nodes (ingress and egress bandwidth). Detailed table with example in figure 4, page 3.

Ingress bandwidth in specific can be reduced by up to 4 times by edge computing. This comes under the umbrella of "Scalability via edge analytics".

Take an example of a Metropolitan City Network with 15,000 users transmitting, downloading, 1080 pixels resolution of video quality, it will consume 110 gigabytes per second, thereby, thoroughly congesting the network, may result in buffering, and at times even an inability to load a video for long period of time. [1]

Edge systems offer a local cache, reduce burden on the entire system, and video streaming would be easily serviced.

C. *In case of node failure*

One may ask why would nodes with good security and management fail? At times it happens due to uncontrollable reasons, which cannot be prevented, the point being it is impossible to uproot the root causes of node failure.

For example, there may be a natural disaster, if there is fire in the building, if there is cyber-attack by hacker, or if a virus has been planted due to a trojan horse, MITM (Man-in-the-Middle), Denial of Service attack, SQL injection, Business email compromise, DNS tunneling, crypto jacking, Drive-by attack, Cross-site scripting attacks and so many more. [9]

Just as recently as May 2017, WannaCry virus had taken over the world and had infected laptops and servers worldwide. Therefore, it is necessary to firstly TRY to prevent a node from failing, but also necessary to have a good backup option, and this is where edge computing had its solution.

In centralized cloud networks, there always is the fear of a single point of failure, this can totally be abolished in a hierarchical network, where distributed dispersed systems are set up, and the cache/local data/essential data can be copied and stored in a system network in layers above in the hierarchy.

This not only prevents data from being totally erased forever in the case of failure of one node, but also this concept can be extended for masking cloud outages.

Masking refers to replicating a nodes (say node A) responsibility by some other node (say node B), in the case the node (node A) is dead. Node B may lie below or above node A in the hierarchy, depending on its congestion, and traffic and capacity, it can either perform only critical operations or it may take over all the responsibilities of the failed node. This way, a single point of failure will no longer be a huge threat.

D. *Security in nodes*

A prevalent problem in today's internet is excessive centralization of user data in cloud servers and data centers.

Due to cookie sharing and company policies, users don't have the control as to what data of theirs is channeled where and to who and its going and to who all this data will be further distributed.

This is a huge and growing concern as users tend to store personal and sensitive data in sites and this may be wrongly intercepted by hackers or just plainly distributed by companies to other companies as part of their policies for commercial reasons.

Either way, users only have one option of accepting or denying cookies, and as some sites do not allow access without accepting cookies, users may not have a choice and have made themselves vulnerable on the internet.

In edge computing, the user can enforce his policies and have full control of what data is sensitive and cannot be leaked and what data can be distributed for constructive purposes. As the edge will be located in his own subnet, he can be in total control of what data can be sent to the main centralized parent cloud server and what the cloudlet can locally cache.

Another advantage of this is in the case of attack to centralized cloud system, the user's sensitive data is not made vulnerable to the attacker as it is completely his own choice to store it there or not. Edge computing is therefore, a huge step up in the cyber security issue faced by the world today.

E. *MEC is the key to 5G*

5G PPP (Public Private Partnership) recognize MEC (mobile edge computing) which has been standardized by the ETSI ISG. 5G networks have a vision to incorporate a programmable interface to networking. [4][6]

Edge computing can help transform it from a broadband network to a programmable one, hence AI applications can be run and since edge computing lies in close proximity to its users.

5G can even meet the needs of bandwidth and latency of applications sensitive to it and require extremely fast transmission, processing and receiving of data.

In addition, to reduce the first hop time, Ultra-Reliable Low-Latency Communication (URLLC) service is implemented in the edge computing network in the 5G framework. This reduce latency by a huge margin and meets the demands on the applications that need it.

III. Conclusion

A. *Current problems faced by Edge Computing*

Edge computing requires setting up of computational, data storage node all across the geographical map, which is costly. In addition, to make sure its safe from being compromised, security applications need to be installed, which is again costly and it won't be of the same scale of security as a big centralized data center present in Companies.

Another issue is improving of links all across the network. Cloud storage systems generally need to have data links of high bandwidth and large capacity near the data center for proper functioning. For edge computing, for each user to have a smooth experience, each data link in network edge will need to be of good quality and this scaling throughout the network is costly. [5]

B. *Future aspects of Edge Computing*

To popularize Edge computing, an open eco system needs to be set up, just as internet was opened up to the world. This

Application Name	Application Description	Crunched metadata
Table Tennis Coach	A VR headset can we worn while playing a game of table tennis and the opponent's head position, body position, previous shots played, probability of success of them will be taken into account. The output would be which direction should the player move and what shot he should play. This can hone a player's skill and intuition as he will intuitively develop which direction to move after playing a shot.	<ball position, speed, position of opponent player, position of opponent players racquet>
Heartbeat	Smart watches can have an app to read ECG reading and spo2 value and send to an edge computing node for Monitoring and store an owner's heartbeat. In case of heart attack, arrange for an ambulance to a hospital closest to the victim. This can be extended to detect arrythmia and other heart related diseases.	<ECG reading, spo2 level, GPS position (user sensitive data, unreleased by cloudlet to central data center until the case of heart attack or any other emergency)>
Accident Prevention	Automobiles can employ this technology paired with a mobile device (Most car owners will have a mobile phone). The speed of the vehicle can be monitored, paired with steering and GPS position along with past experience with the driver and AI mechanisms can sense if driver is drowsy or under the influence of alcohol, and then send an alert to the driver or give a warning. In case of accident, sudden slam of breaks, windows are automatically lowered so that driver and passengers are not engulfed and suffocated in the case of fire.	<GPS position (user sensitive data, unreleased by cloudlet to central data center until the case of heart attack or any other emergency), vehicle speed, steering, brake frequency and pressure, acceleration>
Rubix Cube solver	A VR headset, phone camera can be used to scan unsolved Rubix cube and then give the solution to solve it step by step.	<position of center pieces, coordinates of moveable pieces, changes made>
Encyclopediascope	A VR headset or smart goggles can we worn, and via visual data, famous people can be identified, books, articles and items and their uses and facts can be viewed by the user after recognition.	<compressed (pixels) frames sent and minimal frames sent if objects in the view of lens do not change.>
Google lens and smart other lenses	Smart lens can we worn and if you want to capture a picture all you have to do is blink twice. That way you never miss capturing a moment and is as simplistic as it gets. The picture can be transferred to the phone. Smart lens uses an edge node for this processing.	<blink-flag, timestamp, user data seen through his lens (overwritten, this way the user need not worry for sensitive data to be leaked to data centers as edge computers just transfers the pictures to the phone if taken, if phone is not reachable, cache it until phone is reachable and then delete the data. Cloud servers up the hierarchy are unused.)>

Fig. 4. Conceptual applications which require edge computing for optimal usage and their metadata values which will be the data transferred.

would ensure large participation of companies and people to join the revolution of edge computing.

As discussed before, Edge Computing is essential in the development of 5G Internet. [4][6]

Edge Computing principles ares continuing its progress in wearable new devices such as cognitive assistance goggles, or watches, mobiles(MEC's) and even automobiles.

Edge computing not only delivers data at low latency and high bandwidth, but also takes up all the computation power upon itself and the devices work as simple terminal devices, which give input and receive output, thereby, complex programs can be run in edge computers without any extra battery consumption of low-end devices.

Cognitive assistance goggles are having a huge uprise in research and development as their possibilities are endless. They can be implemented to run face recognition softwares, to navigation softwares, to softwares which can instruct you on how to solve problems like rubix cube's and sudoku's.

Automobiles and Aeroplanes can have edge connected nodes which can run accident prevention mechanisms and turbulence prevention mechanisms respectively.

The field of edge computing basically is a portal to integrate Artificial Intelligence programs to devices which we use in our daily lives. [3]

REFERENCES

[1] M. Satyanarayanan, "The Emergence of Edge Computing," in Computer, vol. 50, no. 1, pp. 30-39, Jan. 2017, doi: 10.1109/MC.2017.9.

[2] W. Shi and S. Dustdar, "The Promise of Edge Computing," in Computer, vol. 49, no. 5, pp. 78-81, May 2016, doi: 10.1109/MC.2016.145.

[3] W. Shi, J. Cao, Q. Zhang, Y. Li and L. Xu, "Edge Computing: Vision and Challenges," in IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637-646, Oct. 2016, doi: 10.1109/JIOT.2016.2579198.

[4] N. Abbas, Y. Zhang, A. Taherkordi and T. Skeie, "Mobile Edge Computing: A Survey," in IEEE Internet of Things Journal, vol. 5, no. 1, pp. 450-465, Feb. 2018, doi: 10.1109/JIOT.2017.2750180.

[5] B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick and D. S. Nikolopoulos, "Challenges and Opportunities in Edge Computing," 2016 IEEE International Conference on Smart Cloud (SmartCloud), 2016, pp. 20-26, doi: 10.1109/SmartCloud.2016.18.

[6] Yun Chao Hu, Milan Patel, Dario Sabella, Nurit Sprecher and Valerie Young, "Mobile Edge Computing A key technology towards 5G," September 2015, ISBN No. 979-10-92620-08-5.

[7] M. Satyanarayanan et al., “The Case for VM-Based Cloudlets in Mobile Computing,” IEEE Pervasive Computing, vol. 8, no. 4, 2009, pp. 14–23.

[8] F. Bonomi et al., “Fog Computing and Its Role in the Internet of Things,” Proc. 1st Edition MCC Workshop Mobile Cloud Computing (MCC 12), 2012, pp. 13–15.

[9] Adleman L.M. (1990) An Abstract Theory of Computer Viruses. In: Goldwasser S. (eds) Advances in Cryptology - CRYPTO' 88. CRYPTO 1988. Lecture Notes in Computer Science, vol 403. Springer, New York, NY.

[10] M. Satyanarayanan, "Pervasive computing: vision and challenges," in IEEE Personal Communications, vol. 8, no. 4, pp. 10-17, Aug. 2001, doi: 10.1109/98.943998.